Skeleton Based Action Recognition with Convolutional Neural Network

Yong Du[†, ♮], Yun Fu[♯], Liang Wang[†, ‡, ♮]
[†]Center for Research on Intelligent Perception and Computing, CRIPAC
[‡]Center for Excellence in Brain Science and Intelligence Technology, CEBSIT
[♮]Nat'l Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[♯]College of Engineering, College of Computer and Information Science, Northeastern University, USA
{yong.du, wangliang}@nlpr.ia.ac.cn, yunfu@ece.neu.edu

Abstract

Temporal dynamics of postures over time is crucial for sequence-based action recognition. Human actions can be represented by the corresponding motions of articulated skeleton. Most of the existing approaches for skeleton based action recognition model the spatial-temporal evolution of actions based on hand-crafted features. As a kind of hierarchically adaptive filter banks, Convolutional Neural Network (CNN) performs well in representation learning. In this paper, we propose an end-to-end hierarchical architecture for skeleton based action recognition with CNN. Firstly, we represent a skeleton sequence as a matrix by concatenating the joint coordinates in each instant and arranging those vector representations in a chronological order. Then the matrix is quantified into an image and normalized to handle the variable-length problem. The final image is fed into a CNN model for feature extraction and recognition. For the specific structure of such images, the simple max-pooling plays an important role on spatial feature selection as well as temporal frequency adjustment, which can obtain more discriminative joint information for different actions and meanwhile address the variable-frequency problem. Experimental results demonstrate that our method achieves the state-of-art performance with high computational efficiency, especially surpassing the existing result by more than 15 percentage on the challenging ChaLearn gesture recognition dataset.

1. Introduction

Action recognition plays an important role in computer vision and has a wide range of applications, *e.g.*, human-computer interaction, video surveillance, robotics, game control, and so on [1, 24]. Generally, human body can be regarded as an articulated system with rigid bones and hinged joints, and human actions can be represented as the motions of skeleton [19]. Currently, the cost-effective depth sensor combining with real-time skeleton estimation algorithms [15, 16] can provide relatively reliable joint coordinates. Based on those coordinates, effective and efficient approaches for action recognition have been developed recently [5, 19, 26].

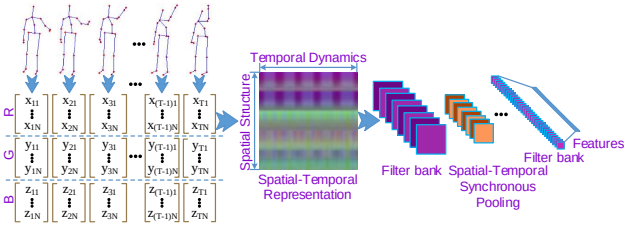

Figure 1: An illustrative sketch of the proposed method. Three components of all skeleton joints in each frame are separately concatenated by their physical connections. After arranging the representations of all frames in chronological order, the generated matrix is quantified and normalized into an image, which is fed into the hierarchical spatial-temporal adaptive filter banks model for representation learning and recognition.

Temporal dynamics of postures over time can be modeled as a time series problem, which is crucial for sequence-based action recognition [4, 8, 9]. As a kind of low-level feature, skeleton joint coordinates can be used to represent human postures and their temporal evolution. Most of the existing skeleton based action recognition approaches model actions based on well-designed hand-crafted local features. Simultaneously, Temporal Pyramids (TPs) and its variants are often employed to capture the local temporal evolution [10, 19, 20]. A following dictionary learning model is employed to generate the representation for the whole sequence [20]. For the restriction from the width of time windows, the TPs methods can only utilize limited contextual information. Moreover, temporal dynamics of

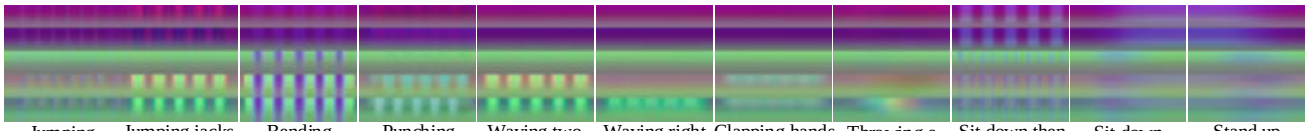

Figure 2: Image representations obtained on the Berkeley MHAD dataset [12].

sequences is ignored in the global representation based on a learned dictionary. Alternatively, some time series models, mainly Dynamic Time Warpings (DTWs) [19] and Hidden Markov Models (HMMs) [11, 21, 23], are applied to model the global temporal evolution, yet it is very difficult to obtain the temporal aligned sequences and the emission distributions of HMMs. Recently, an end-to-end approach based on Recurrent Neural Network (RNN) was proposed for skeleton based action recognition [1].

Almost all the existing methods for skeleton based action recognition concern that how to utilize the contextual information to capture temporal dynamics in sequences. HMMs and RNNs based approaches directly model the temporal dynamics as a time series problem. However, the temporal dynamics of sequences can be easily transformed as the spatial structure characteristics in images. As an architecture with hierarchical adaptive 2D filter banks, Convolutional Neural Network (CNN) has the advantage of encoding structural information and can be used for representation learning in many tasks. In this paper, we propose a CNN model for skeleton based action recognition. Firstly, we represent each skeleton sequence as a special image, in which the temporal dynamics of the sequence is encoded as changes in rows and the spatial structure of each frame is represented as columns (Fig. 1). A following CNN model is employed for feature learning and recognition. To represent the skeleton structure compactly, we employ the encoding style in [1] that human skeleton is divided into five parts and joints in each part are concatenated according to their physical connections. And three components (x, y, z) of each joint are represented as the corresponding three components (R, G, B) of each pixel. Considering that the length of skeleton sequences are variable, we resize the generated images to an uniform size yet the frequency of actions is changed with different scales. So spatial-temporal synchronous pooling is used to overcome this variable-frequency problem. We evaluate our approach on two benchmark datasets and obtain excellent performance.

The main contributions of our work can be summarized as follows. Firstly, we propose an idea to represent action sequences as images as well as preserving the original temporal dynamics and spatial structure information. Secondly, based on such representation, we propose a CNN model for skeleton based action recognition, in which the sample 2D max-pooling plays the role of spatial-temporal synchronous pooling and can overcome the different sequence length and variable-frequency problems. Finally, we demonstrate that our proposed end-to-end model can rapidly handle skeleton based action recognition very well without any sophisticated processing. And this idea can be easily transfered to other time series problems.

The remainder of this paper is organized as follows. In Section 2, we introduce our proposed model in detail. Then we provide our experimental results in Section 3. Finally, we conclude the paper in Section 4.

2. Our Method

In this section, we first detail how to represent a sequence as an image, and then we introduce our hierarchical model for skeleton based action recognition. Finally, more details about training and testing are provided.

2.1. From Skeleton Sequences to Images

How to transform a sequence to an image while preserving its spatial-temporal information is very important. Inspired by [1], all human skeleton joints in each frame are divided into five main parts according to human physical structure, *i.e.*, two arms, two legs and a trunk. For preserving the local motion characteristics, joints in each part are concatenated as a vector by their physical connections, *e.g.*, each arm can be represented as [hand, wrist, elbow, shoulder]. Then the five parts are concatenated as the representation of each frame. Projections on three orthogonal planes are represented separately and treated as the three components of RGB images, *i.e.*, $R_i = [x_{i1}, x_{i2}, \ldots, x_{iN}]$, $G_i = [y_{i1}, y_{i2}, \ldots, y_{iN}]$, $B_i = [z_{i1}, z_{i2}, \ldots, z_{iN}]$, where i denotes the frame index and N indicates the number of frames in a sequence. Finally, representations of all frames are arranged in chronological order to represent the whole sequence (Fig. 2). In this case, spatial distribution of motion characteristics of each part is very clear, and the global discrimination is pretty obvious. Given the variable-length problem, the arranged float matrix is quantified to integral image representation, *i.e.*,

$$p = floor\Big(255 * \frac{p - c_{min}}{c_{max} - c_{min}}\Big) \tag{1}$$

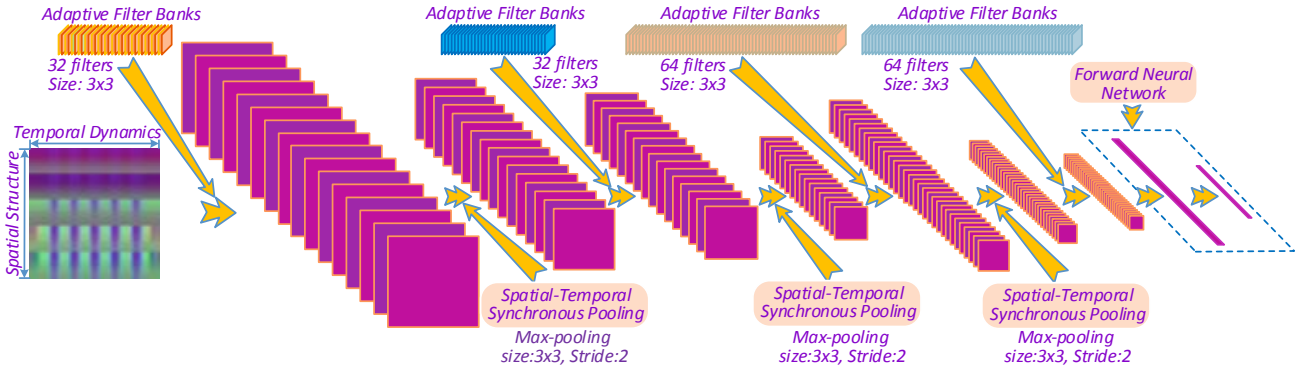

Figure 3: The framework of our model based on hierarchical spatial-temporal adaptive filter banks.

then the image is resized to an uniform size 60×60. The p indicates the pixel value and the floor function means rounding down. The c_{max} and c_{min} are the maximum and minimum of all joint coordinates in the training set, respectively.

2.2. Hierarchical Architecture for Skeleton Based Action Recognition

Recognizing actions depends on not only human pose in each time but also its temporal dynamics [1, 4, 8]. Generally, an adaptive filter bank can learn the filter coefficients adaptively and separate the input signal into multiple components. And those components can be analyzed in different domains corresponding to the diverse filter characteristics. Given that the essence of CNN is composed of a range of adaptive 2D filter banks, different from traditional HMMs or other time series models based approaches directly modelling action recognition as a time series problem, we propose a CNN based model to capture the spatial-temporal information in the sequences after transforming them to special image representations.

Our proposed model is shown in Fig. 3, which includes 4 cascaded adaptive filter banks. All filter sizes are 3×3 and all strides during convolution are set to 1 for treating CNN as hierarchical adaptive filter banks. Considering that the original action frequencies are changed in different scales when resizing and same actions performed by different subjects may have various frequencies, we adopt the max-pooling strategy following each of the first three filter banks. For the special structure of input images (Fig. 2), the scale-invariance of max-pooling along horizontal axis is transformed as the frequency-invariance of actions. And max-pooling along vertical axis can select more discriminative skeleton joints for different actions. After feature extraction, a feed-forward neural network with two fully-connected layers is employed for classification. The first fully-connected layer contains 128 neurons, and the number of neurons in the second one is equal to that of actions. ReLU neuron is adopted in all layers. The loss function is as follows [1]:

$$\mathcal{L}(X) = -\sum_{m=0}^{M-1} \ln \sum_{k=0}^{C-1} \delta(k-r) p(C_k|x_m) \qquad (2)$$

where M indicates the number of samples in training set X and C denotes the number of categories. $\delta(\cdot)$ is the Kronecker function, and $p(C_k|x_m)$ means the probability that sample x_m belongs to action C_k.

2.3. Training and Testing

Our model is trained by the back propagation algorithm. In order to improve the convergence speed, mean removal is employed for all input images. During training, we randomly select a patch from each original image with size 52×52. Given that temporal dynamics of a skeleton sequence is always embodied in two directions, we randomly flip the training images horizontally to utilize the forward and backward temporal dynamics. During testing, five patches selected form the four corners and the center and their horizontal flips are used. The final recognition is obtained by voting.

3. Experiments

In this section, we evaluate our model compared with several recent works on two benchmark datasets: Berkeley Multimodal Human Action Dataset (Berkeley MHAD) [12] and ChaLearn gesture recognition dataset [2]. We also discuss the computational efficiency of our model.

3.1. Evaluation Dataset

Berkeley MHAD [12]: It is generated by a multimodal acquisition system and an optical motion capture system is employed to capture the 3D position of skeleton joints with the frequency of 480Hz. There are 659 valid samples in this dataset, which consists of 11 actions performed by 12 subjects with 5 repetitions of each action. And each frame in a

sequence contains 35 joints accurately extracted according to the 3D marker trajectory.

ChaLearn Gesture Recognition Dataset [2]: It is the ChaLearn 2013 Multi-model gesture dataset, which contains 23 hours of Kinect data with 27 persons performing 20 Italian gestures. This dataset provides RGB, depth, foreground segmentation and Kinect skeletons. This dataset is split into training, validation and testing sets, and contains total 955 videos, each of which lasts 1-2 minutes and involves 8-20 non-continuous gestures.

3.2. Experimental Results and Analysis

Berkeley MHAD: We follow the experimental protocol proposed in [12]. The 384 sequences of the first 7 subjects are used for training while the 275 sequences of the last 5 subjects are used for testing. We compare our proposed approach with Ofli *et al*. [13], Vantigodi *et al*. [17], Vantigodi *et al*. [18], Kapsouras *et al*. [6] and Du *et al*. [1]. All the comparative methods on this dataset are directly from their corresponding papers, and the rest likewise. The experimental results are shown in Tab. 1. We can see that our method can achieve the 100% accuracy without any other pre- or post-processing. In contrast to those handcrafted features based approaches [6, 13, 17, 18], our approach, like [1], is an effective end-to-end solution to skeleton based action recognition.

Table 1: Experimental results on the Berkeley MHAD [12].

Method	Accuracy (%)
Ofli *et al*., 2014 [13]	95.37
Vantigodi *et al*., 2013 [17]	96.06
Vantigodi *et al*., 2014 [18]	97.58
Kapsouras *et al*., 2014 [6]	98.18
Du *et al*., 2015 [1]	100
Ours	**100**

ChaLearn Gesture Recognition Dataset: In this more challenging dataset, the ground segments are provided and contain 6850 training samples with 39 frames average length, 3454 validation samples and 3579 test samples. We follow the experimental protocol adopted in [3, 14, 22, 25] and provide precision, recall and F1-score measures on the validation set. We compare our model with Yao *et al*. [25], Wu *et al*. [22], Pfister *et al*. [14], and Fernando *et al*. [3]. The experimental results are shown in Tab. 2. It is clear that our method significantly surpass the state-of-the-art precision by more than 15 percentage, which demonstrate that it is a great success to transform temporal dynamics in sequences into spatial structure information in images for sequence representation learning. One of the possible reasons for the excellent performance may be that our model can well handle the global temporal dynamics in sequences than the comparative methods. And comparing with those recently proposed approaches, our method is simple and straightforward for skeleton based action recognition. Filters and the convergence curves are shown in Fig. 4.

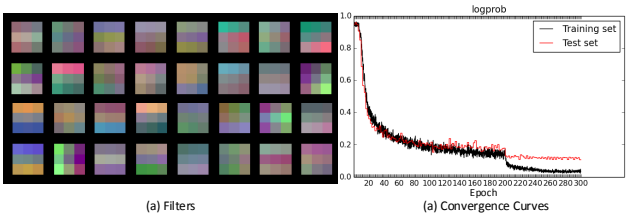

Figure 4: Filters and convergence curves on the ChaLearn gesture recognition dataset.

Table 2: Experimental results on the ChaLearn gesture recognition dataset [2].

Method	Precision	Recall	F1-score
Yao *et al*., 2014 [25]	-	-	56.0
Wu *et al*., 2013 [22]	59.9	59.3	59.6
Pfister *et al*., 2014 [14]	61.2	62.3	61.7
Fernando *et al*., 2015 [3]	75.3	75.1	75.2
Ours	**91.16**	**91.25**	**91.21**

3.3. Computational Efficiency Analysis

We take the ChaLearn gesture recognition dataset for an example to illustrate the efficiency of our proposed model implemented based on ConvNet [7]. With the implementation on NVIDIA Titan GK110, we spend about 1.95ms per sequence in training and about 2.27ms per sequence (select 5 patches and flip for voting) in testing.

4. Conclusion and Future Work

In this paper, we have proposed a simple end-to-end but high-efficiency and high-precision framework for skeleton based action recognition. We first represented human skeleton sequences as images to transform the temporal dynamics of sequences into the spatial structure information in images. Then a hierarchical architecture based on CNN was proposed for feature representation learning and classification. Experimental results on two publicly available datasets demonstrated the excellent performance of the proposed model.

Current model classifies actions based on the global spatial and temporal information in skeleton sequences, which requires that the noise distribution in different segments of the same sequence are consistent. That means if data error of local fragments in the input sequences is particularly highlighted, the recognition rate may be cut down. In the future, we will consider the local features as an assistance to overcome this problem.

Acknowledgement

This work is jointly supported by National Natural Science Foundation of China (61420106015, 61175003) and National Basic Research Program of China (2012CB316300). Yun Fu is supported by National Science Foundation (NSF) CNS award 1314484.

References

[1] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1110–1118, 2015. 1, 2, 3, 4

[2] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *ACM on International Conference on Multimodal Interaction*, pages 445–452. ACM, 2013. 3, 4

[3] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars. Modeling video evolution for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5378–5387, 2015. 4

[4] D. Gong, G. Medioni, and X. Zhao. Structured time series analysis for human action segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 36, NO. 7*, 2014. 1, 3

[5] C. Jia, G. Zhang, and Y. Fu. Low-rank tensor learning with discriminant analysis for action classification and image recovery. In *AAAI Conference on Artificial Intelligence*, pages 1228–1234, 2014. 1

[6] I. Kapsouras and N. Nikolaidis. Action recognition on motion capture data using a dynemes and forward differences representation. *Journal of Visual Communication and Image Representation*, 25(6):1432–1445, 2014. 4

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 4

[8] K. Li and Y. Fu. Prediction of human activity by discovering temporal sequence patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 36, NO. 8*, 2014. 1, 3

[9] K. Li, J. Hu, and Y. Fu. Modeling complex temporal composition of actionlets for activity prediction. In *European Conference on Computer Vision*, pages 286–299, 2012. 1

[10] J. Luo, W. Wang, and H. Qi. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *IEEE International Conference on Computer Vision*, pages 1809–1816. IEEE, 2013. 1

[11] F. Lv and R. Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *European Conference on Computer Vision*, pages 359–372. Springer, 2006. 2

[12] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. In *IEEE Workshop on Applications of Computer Vision*, pages 53–60. IEEE, 2013. 2, 3, 4

[13] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation*, 25(1):24–38, 2014. 4

[14] T. Pfister, J. Charles, and A. Zisserman. Domain-adaptive discriminative one-shot learning of gestures. In *European Conference on Computer Vision*, pages 814–829. Springer, 2014. 4

[15] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013. 1

[16] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems*, pages 1799–1807, 2014. 1

[17] S. Vantigodi and R. V. Babu. Real-time human action recognition from motion capture data. In *National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, pages 1–4. IEEE, 2013. 4

[18] S. Vantigodi and V. B. Radhakrishnan. Action recognition from motion capture data using meta-cognitive rbf network classifier. In *International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, pages 1–6. IEEE, 2014. 4

[19] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595. IEEE, 2014. 1, 2

[20] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1290–1297. IEEE, 2012. 1

[21] D. Wu and L. Shao. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In *IEEE International Conference on Computer Vision*, 2014. 2

[22] J. Wu, J. Cheng, C. Zhao, and H. Lu. Fusing multimodal features for gesture recognition. In *ACM on International Conference on Multimodal Interaction*, pages 453–460. ACM, 2013. 4

[23] L. Xia, C.-C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–27. IEEE, 2012. 2

[24] X. Yang and Y. Tian. Super normal vector for activity recognition using depth sequences. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 804–811. IEEE, 2014. 1

[25] A. Yao, L. Van Gool, and P. Kohli. Gesture recognition portfolios for personalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1923–1930. IEEE, 2014. 4

[26] X. Zhao, Y. Liu, and Y. Fu. Exploring discriminative pose sub-patterns for effective action classification. In *ACM International Conference on Multimedia*, pages 273–282, 2013. 1